ON DETECTING TRANSIENT PHENOMENA

G. Bélanger[1]

[1] European Space Astronomy Centre (ESA/ESAC), Science Operations Department, Villanueva de la Cañada (Madrid), Spain

ABSTRACT

Transient phenomena are interesting and potentially highly revealing of details about the processes under observation and study that could otherwise go unnoticed. It is therefore important to maximise the sensitivity of the method used to identify such events. In this article we present a general procedure based on the use of the likelihood function for identifying transients that is particularly suited for real-time applications because it requires no grouping or pre-processing of the data. The method makes use of all the information that is available in the data throughout the statistical decision making process, and is suitable for a wide range of applications. We here consider those most common in astrophysics which involve searching for transient sources, events or features in images, time series, energy spectra and power spectra, and demonstrate the use of the method in the case of a weak X-ray flare in a time series and a short-lived QPO in a power spectrum. We derive a fit statistic that is ideal for fitting arbitrarily shaped models to a power density distribution, which is of general interest in all applications involving periodogram analysis.

Subject headings: methods: data analysis — methods: statistical

1. INTRODUCTION

Some physical processes can be considered continuous in the sense that the discretisation of measurements comes from the detector's sampling rate. Others are discrete in the sense that they give rise to individual events that can be detected as such, as long as the sampling is faster than the detection rate. For non-variable processes, the values of the measurements will generally be distributed as either a Normal variable—in those seen as continuous, or as a Poisson variable—in those that are discrete.

The way in which the measurements are distributed is absolutely crucial in applying the appropriate statistical treatment. But irrespective of that distribution, each measurement carries information that can be used to estimate the values of the parameters of models that help us learn about the processes being observed. Most importantly, each measurement considered individually, and the collection of measurements as a whole, all carry statistical evidence that can be used to accurately assess the agreement between a given hypothesis or model and the data.

Treating data as evidence is a powerful means to detect changes, differences, deviations or variations in any kind of process that is observed. Treating data as statistical evidence is, in fact, the only way that data should be treated no matter what the application or circumstances, because that is what data actually is. And the way this is done, mathematically and statistically, is through the likelihood function.

As obvious as this is, it is important to point out that the detection of an event or feature that is localised in time, involves identifying something that was not there before. Whether it rises, dwells, and decays over weeks and months like a supernova, or whether it just appears and disappears in a fraction of a second like a γ-ray burst; whether it manifests as a complete change of shape of the energy spectrum during a state transition in a black hole, or as the short-lived emission line from an accretion event; whether it comes as a sudden change of spectral index in the power spectrum or as the appearance of an ephemeral quasi periodic oscillation (or QPO); all of these phenomena, independently of their particular time scale, share

in common that they appear as a sharp change in the data.

Detection and identification of a transient feature in an ensemble of measurements is a statistical procedure that involves comparing numbers and making decisions based on probabilities or, in fact, on probability ratios, and in other words, on likelihoods. Naturally, we would like to use the most powerful method for the problem at hand. Thus, whatever the application, we want to maximise sensitivity to transient phenomena, and minimise the frequency of identifying a statistical fluctuation as an actual event. For this reason we must use all the information that is available, without degrading in any way the data the instruments provide us with, and interpret these as statistical evidence.

In this article we address this task of identifying *transients*, in the most general sense of the word, without reference to the kind of phenomenon nor the kind of data we are working with. In light of this, we use the word *transient* in a sense that is broader than its customary usage, which refers to a source whose intensity varies dramatically enough to go from being invisible to detectable or even bright, and back to being undetectable. We define a *transient* as any feature or change that can be identified in the data as *statistically distinct* from the global process and underlying conditions. This definition, therefore, implies that if a feature cannot be distinguished by statistical means, it cannot be detected and identified. Whether this is because the transient is too weak or too long-lasting does not matter, and there is hence no need to talk about it: the limitations of a particular detection procedure, with its own thresholds and timescales, can always be accurately established before applying it.

We develop and present a general method based on the likelihood function that can be applied to wide range of problems (§2). We describe the procedure (§2.1), what the likelihood function is (§2.2), and illustrate the method for a general counting experiment (§2.3). We elaborate on the procedure's use and applicability in astronomy and astrophysics (§3) by considering, after a few introductory remarks (§3.1), its application to images (§3.2), time series (§3.3), energy spectra (§3.4) and power spectra (§3.5). We look at the identification

of transients over a variable background in the next section (§2) and in the concluding remarks (§4).

The mathematical statistics of likelihood are from the work of Fisher (1912, 1922); the philosophical basis for, understanding of, and inspiration to work with and interpret data as statistical evidence are primarily from Royall (1997); and other technical details of data analysis and statistics are mostly from Cowan (1997).

2. IDENTIFYING TRANSIENTS

A transient can only be identified as such in relation to something else: in relation to the underlying background conditions. There are two families of circumstances pertaining to the characteristics of the background process that cover all cases under which transients may appear: the process can be constant or it can be variable. In absolute terms, it could be said that if a process is not constant, then it is variable. In practice, however, the variability is characterised by time scales above which the process is seen to be variable, but below which it cannot, or at least not easily be seen to be variable. Furthermore, the variability will in general manifest differently at different time scales.

In most applications where transient sources are searched for, they are of the kind not detectable until they brighten for a while before once more fading away below the detection level. Therefore, the background is characterised by the statistical and instrumental fluctuations inherent to the particular experimental setup and sky pixel where the source appears. This is also generally true for supernovae and γ-ray bursts at all wavelengths (but on different time scales), as well as for X-ray novae or X-ray flares, bursts or flashes in most systems: the underlying background is usually constant or very weakly variable in comparison to the sharp change in intensity associated with the transient phenomenon.[1] In the energy and power spectra, irrespective of spectral shape, transient phenomena will also most often appear against a constant or very weakly variable background. Therefore, in all these cases and in the majority of applications searching for transients, the background is constant or nearly so.

In fact, this is indeed what allows these short-lived phenomena to be considered and labelled as transient. But as the intensity of the background against which we seek to identify a transient of a given magnitude increases, the ability to do so decreases. If instead of increasing in intensity, the background were to increase in its variability, the ability to identify a transient similarly decreases. In both cases, it is a question of scales: of the intensity scale and of the time scale. In the first case, statistical identification depends mostly on the ratio of intensities between the transient and the background (the commonly held notion of signal-to-noise), but also on its duration: the brighter, the easier to identify; and if it is not bright, the longer it lasts, the easier to identify. In the second, identification depends more on the ratio of the transient's duration to the time scale of the variability of the background, but obviously also on the magnitude of its intensity: the shorter in duration, the easier to identify with respect to a slowly variable background; and naturally the brighter it is, the easier the identification.

Hence, these factors both come into play, and gain or lose in importance depending on the characteristics of the physical process being observed, and most sensitively, on the quality of the data. It is important to underline that these elements— intensity ratios and variability time scales—can and should be considered and worked with as parameters in order to establish optimal detection algorithms and thresholds for different problems in various kinds of data sets, as well as to determine the limitations of our methods that should ultimately always be statistical in nature, and dependent on the quality and kind of data; not on weaknesses in the methods themselves. We will carry out and present such an investigation and quantitative characterisation of these issues, including the discussion of commonly encountered complications in a future work, together with applications to data from different experiments and surveys.

The purpose of this paper, is to present the foundational aspects of the method, and illustrate how it can be of immediate applicability in a variety of situations that include the search for transient features in X-ray and γ-ray data, transient sources in optical sky surveys, radio transients and in particular RRATs that are currently typically identified by eye. Our basic working assumption, which does indeed cover most applications, is that we are dealing with a background that is constant, (nil or not), on the time scales relevant to the problem of identifying transients.

The method is straight forward and based on the likelihood function. As such, all of the details that pertain to the inherent statistical properties of the kind of random variable that results from the observational process (e.g., Normal, Poisson, Exponential, χ^2) are automatically and effortlessly taken into account at every step, and integrated in every aspect of the procedure. The presentation is intended to be as clear, intuitive and easy and as possible, with minimal formalism. It is worth noting that the tools for this have been available in practice at least since Fisher (1922), and in principle since Bayes and Price (1763). The method makes use of every measurement, does not require grouping or approximations, and therefore does not impose on the data any degradation of its resolution or accuracy, no matter what kind of data it is. Since the approach is the same in all circumstances where the aim is to detect a transient, it is described in general terms and its principles demonstrated in the next section. Astrophysics applications are discussed afterwards.

2.1. Methodology

The first measurement gives the first estimate of the reference value—the value we expect to measure under usual conditions when there is no transient. With this single measurement we can already draw the curve that expresses the likelihood of all possible reference values given the evidence carried by that measurement.[2] To draw this curve, we must make an informed assumption as to *how the measurements will be distributed* around the true value: we must make an informed assumption about the nature of that random variable. As was mentioned, the most common in physical experiments are the Normal distribution seen in continuous processes like measuring temperature, and the Poisson distribution that arises

---

[1] The work of Scargle (1998) and Scargle et al. (2013) on astronomical time series, with which we were not acquainted while developing our method, is different in its details but quite similar in spirit, (even if more complicated in presentation and implementation), and seems well suited for GRB searches in X-ray and γ-rays time series.

[2] In practice, most observations (from most patches of sky and most sources) are not the first of their kind, and there is, therefore, no need to make a guess of the expected intensity; it can just be determined from previous observations, which implies that even the first measurement can be compared to the expected reference value, and the sensitivity of the method does not depend on the number of prior measurements.

when discrete events are recorded. The likelihood function shows the maximum likelihood of the true value, and the ratio between that and every other possible value: it gives us the relative likelihood of any value with respect to any other, depending only on the nature of the variable and on the data.

The second measurement gives a second estimate of the reference value. Because we already have a likelihood function based on the first measurement, the value of the second can be immediately evaluated for its potential of being a transient—a feature that stands out from what is expected—by computing its likelihood ratio with respect to the previously estimated maximum likelihood. Although limited by the accuracy with which the mean is known, this is nonetheless the only mathematically correct way to evaluate the likelihood of measuring that second value in light of the first, without invoking an a priori model or assumption. If the second measurement is not statistically different from the first beyond the established threshold, it is combined with the first to better estimate the mean, and can also be used to compute a first estimate of the standard deviation of the distribution. The joint likelihood function is computed from the two measurements and its central peak immediately begins to grow narrower and hone in on the true value of the parameter.

The third measurement gives a third estimate of the reference, the likelihood of measuring such a value is evaluated by the ratio of the single-measurement likelihood function centred on the maximum likelihood reference value given by the previously calculated joint likelihood. This is an important point: the joint likelihood is the likelihood function of the reference value given the entire set of measurements considered together as a whole, and with each additional measurement, it gets narrower and more finely peaked on the maximum likelihood reference value; the single-measurement likelihood is the function that shows how likely it is to measure any given value each time a measurement is made. The more precisely the reference value is determined, the more reliable the location of the single-measurement likelihood function. However, its shape depends only on the probability density of the random variable and on the reference value.[3]

With each subsequent measurement, the same procedure is repeated: 1) compute the likelihood of the newly measured value based on the single-measurement function defined by the current maximum likelihood reference value; 2) if the likelihood is less than the defined threshold, issue a transient event trigger. Do not update the estimate of the reference value; 3) if the likelihood is more than the defined threshold (within the likelihood interval), recalculate the joint likelihood function including the new measurement and update the maximum likelihood reference value. The single or multiple thresholds used to identify the occurrence of a transient event must be defined and optimised according to the application.

2.2. *The likelihood function*

---

[3] We can formally incorporate the uncertainty in the determination of the reference value into the single-measurement likelihood function by computing the cross-correlation of the joint and single-measurement functions, an operation which yields a broadened function that tends to the pure probability density as the joint likelihood tends towards a delta function. We have investigated this, and found that it increases the complexity of the procedure substantially, but that the effect is only appreciable in the first few measurements where a broadening is seen. Beyond even a handful of measurements, the joint likelihood function is narrow enough for the broadening to be negligible. It is therefore not warranted to include this step unless we are dealing with a process that will only ever yield a handful of measurements.

The likelihood is proportional to the product of probabilities. But because there are an infinite number of points on the probability density function, the product of a collection of probabilities will in general be unbounded. Hence, only the ratio of likelihoods is meaningful. Although it does resemble in shape a probability distribution, it is not: only its shape matters and not its scale. In the words of Fisher (1922, p. 327), "the likelihood may be held to measure the degree of our rational belief in a conclusion", and in those of Royall (1997, p. 28), "Probabilities measure uncertainty and likelihood ratios measure evidence." It is worth emphasising this point: before making a measurement, we have probabilities; after making the measurement, we have likelihoods. We adopt the convenient approach suggested by Fisher (1922) himself, and normalise the likelihood function such that the maximum equals one. This then implies that the value of every point on the curve is already the ratio to the maximum likelihood.

A fundamental distinction is that probability density relates to the random variable whereas likelihood relates to the parameter. The probability density function expresses how we can expect the measured values of the random variable to be distributed given a certain value of the parameter. For example, if we take the rate parameter ν of the Poisson distribution to be equal to 5 events per second, the density function tells us that the probability to detect 5 events in a one second interval is given by the value of the density function at 5 and equals 14.6%, or that the probability to detect 10 events or more is given by the sum from 10 onwards and equals 1.4%.

Now, knowing that we are observing a Poisson process, say that we measure 5 events in the sampling time of one second. That measurement is made and the uncertainty about its value therefore disappears. The value is now evidence about the probability distribution, about the unknown value of the parameter. The likelihood function represents this evidence: it tells us that the most likely value of the rate parameter ν is 5, and that, for example, it is 1.1 times more likely than 6, 4.6 times more likely than 10, and 32 times more likely than 13.4. Computing the likelihood function, the *y*-axis is the relative likelihood and the *x*-axis represents different values of the parameter. In the case of a single measurement n_0, the likelihood associated with each value of the parameter ν on the abscissa is given by $f(n_0; \nu)$; for two measurements n_1 and n_2, it is given by the product $f(n_1; \nu)f(n_2; \nu)$.

In this work we consider four families of random variables: the Poisson, Normal, χ^2 and Exponential. For simplicity and clarity of presentation, the relevant equations are given in Appendix A instead of being presented or derived in the text.

2.3. *Illustration of Method*

The instrument is turned on and the observation begins. Nothing is known other than that we are observing a non-variable Poisson process. The sampling rate is one second and in the first interval three events are detected. With this first measurement, we can already compute the likelihood function of the rate parameter ν, and because we have only one measurement, the joint likelihood is identical to the single-measurement likelihood (Fig. 1, panel 1).

In the second interval, 4 events are detected. We calculate the likelihood ratio for this measurement in relation to the previously estimated maximum likelihood value of the rate parameter which was 3, and find that it is $L_1(\nu = 4)/L_1(\nu = 3) = 0.872$ which is much larger than the warning threshold of $1/8 = 0.125$. Therefore, we compute the joint likelihood function using both measurements, and now have a slightly more
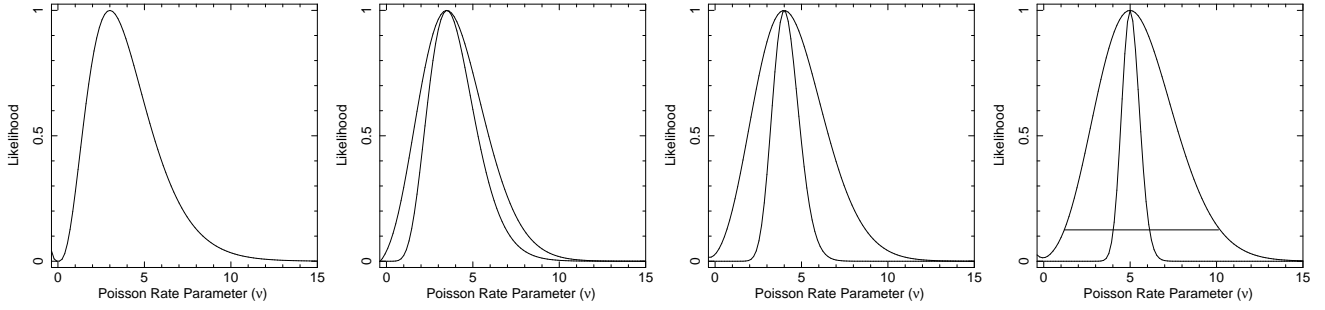
Figure 1. Likelihood functions of a Poisson variable
Graphical representation of the single-measurement and joint likelihood functions after one (panel 1), two (panel 2), seven (panel 3) and nineteen (panel 4) measurements of a Poisson variable with a true rate parameter of $\nu = 5$. The 1/8 likelihood interval is shown in panel 4.

accurate estimate of the rate parameter as exhibited by the narrower peak of the joint likelihood; the single-measurement function is also updated accordingly (Fig. 1, panel 2).

The observation continues and by the time we have made 7 measurements the joint likelihood is quite a bit narrower (Fig. 1, panel 3), and after 19 measurement, the peak is significantly narrower and peaks on the true rate parameter, $\nu_t = 5$ (Fig. 1, panel 4). Naturally, the more measurements are made, the sharper the peak grows, and hence the accuracy of our maximum likelihood estimate of the rate parameter which in turn defines the single-measurement likelihood function against which we evaluate the evidence for the occurrence of a transient with each new measurement by calculating the likelihood ratio; the 1/8 likelihood interval is drawn and seen to extend from 1.2 to 10.2 in panel 4.

3. ASTROPHYSICAL TRANSIENTS

3.1. *Introductory Remarks*

The bounds between which changes can be detected are defined by the instrument's sampling rate for the shortest timescales, and by the time spanned by the data on the longest: if changes in the system take place much faster than the sampling rate (millisecond pulses sampled on second timescales), or much slower than the duration of the observation (an outburst that lasts longer than the time during which it was observed without appreciable change in brightness), they will not be readily detectable. Naturally, transient searches only have meaning within these bounds.

Within these limits, the granularity in time of each iteration of the likelihood procedure is a crucial factor. If made fine enough compared to the time scale of the transient, the change will be gradual and smooth enough to be unrecognised as such. Therefore, the time between iterations must be chosen to ensure optimal sensitivity to time scales relevant to the problem. If there are several, the solution is to run multiple procedures in parallel, each with a different characteristic time scale. Another solution is to monitor the evolution of the average value of the likelihood ratio as a function of time. This technique relies on the stability of the value of the likelihood ratio, and is as sensitive as our knowledge of the background conditions against which we want to detect transient events, (which grows with each additional measurement), and most importantly on the probability distribution of the measurements. Naturally, we can look at the evolution of the likelihood ratio in each channel, in the joint function or in both to maximise our sensitivity to specific kinds of transients.

3.2. *Transients in Images*

Imaging data is acquired by different means at different wavelengths, but in what concerns the task of identifying transient features in these images, the main requirement is that there must be at least more than two and preferably a collection of images of the same region of the sky. So independently of the actual integration time of each snapshot or the means by which this integration time is chosen, we can treat the problem of having sky pixels, each with an ensemble of measured values that we take as an intensity regardless of the units.

In regards to the intensity, there are two cases: 1) the intensity in every pixel is independent of (not correlated with) the intensity in any other pixel, or 2) the intensity in a given pixel is related to the intensity of neighbouring pixels. If the instrument's point spread function (PSF) is mostly contained within a single detector pixel, then we consider each pixel to give independent intensity measurements and also define the size of independent sky pixels. If this is not the case and the PSF spreads on several detector pixels, then we can either make sky pixels as large as necessary to include a large enough fraction of the PSF in order to be considered independent of neighbouring sky pixels, or we must include the model of the PSF in the analysis.

If pixel intensities are independent of one another, this makes an image a collection of intensity measurements, one per pixel, that each corresponds to a distinct region of the sky. Each snapshot yields one such intensity measurement for each pixel, and therefore the problem immediately reduces to the illustrative case above: with the first snapshot, we already have the means to construct the model for each pixel of what can be expected; with each additional snapshot, the model improves in accuracy, and consequently, our ability to detect changes in intensity increases; and the expected intensity in each sky pixel is represented by the single-measurement likelihood function centred and scaled in accord with the joint likelihood function constructed from the ensemble of intensity measurements.

The two relevant functional forms are those of the Poisson and the Normal density distributions (Eqs. (A1) and (A6)). The procedure was illustrated in §2.3 using the Poisson function (Eqs. (A2) and (A3)), but maybe in most imaging applications, the Normal likelihood function (Eqs. (A7) and (A8)) will be the appropriate one to use.

If the detector pixels are smaller than the PSF and the intensity measurements are hence correlated with neighbouring ones, the problem is more complex in its details but the same in its principle. The most significant difference in the approach is that instead of treating images as a collection of independent intensity measurements, they must be considered
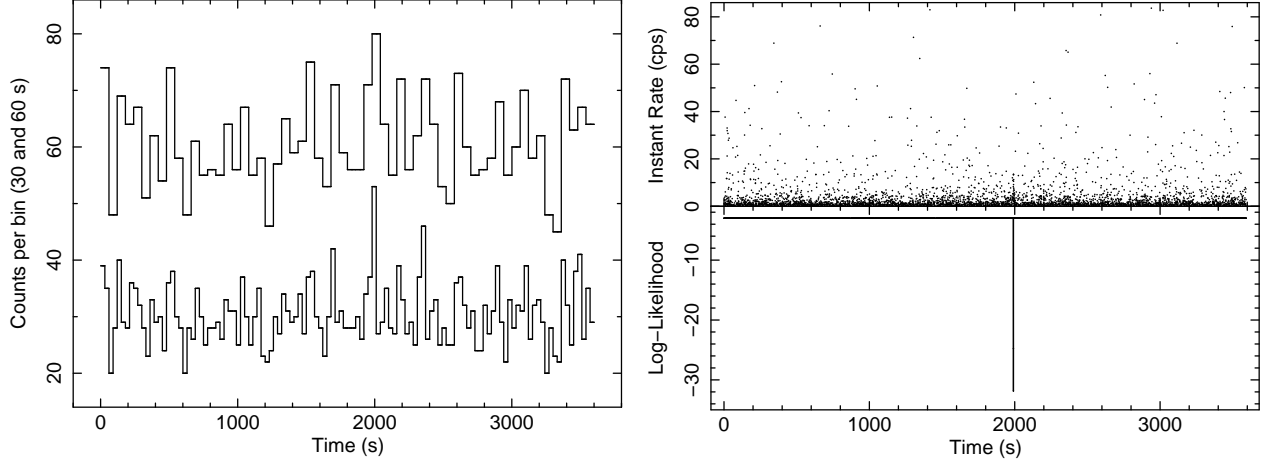
Figure 2. Detection of a transient event in a time series
Time series of the observation shown in counts per bin for 30 and 60 second bins (left panel, bottom and top respectively). Instantaneous count rate calculated as the inverse of the time between events shown as a function of each event's arrival time above the transient detection likelihood also evaluated in real time. The maximum value of the instantaneous rate is 2950 counts per second (cps), but the scale was truncated to 86 cps to match the scale of the 60 s time series and better show the scatter. Values of the log-likelihood that do not meet the trigger criterion are shown at the warning threshold level of -3 (likelihood of 0.05). The sole detection is that of the transient event, and it dips down to -31.85 (likelihood of 10^{-14}).

and modelled as a whole. The global model will preferably be constructed before the observation to include all known sources in the field. With the first image, an intensity is estimated for each identified source included in the model taking into account the instrument's PSF, and then monitored using each subsequent snapshot. Accordingly, each source has its own joint likelihood function that is better defined with each additional image, as well as its own single-measurement likelihood function based on the joint likelihood. These likelihood functions, however, are based on the model of the PSF, and thus only directly on the analytical probability density.

In addition, to the iteratively refined modelling of likelihood functions for each identified source, every other pixel in the image is monitored as if it were independent of all others in the same way as in the previous case. This is obviously of prime importance given that our aim is to detect transients and especially new, weak, and invisible or otherwise unidentified sources. When the intensity of one of these pixels is seen to climb or fall out of its likelihood interval in a number of consecutive snapshots, the new source is then properly modelled using the PSF as for all other sources in the field. The detailed treatment depends on the PSF and is not carried out here for a hypothetical case, but the use of a global model is analogous to the treatment of both energy and power spectra. We thus leave out of this section a more explicit discussion.

3.3. *Transients in Time Series*

The procedure for treating time series is, in fact, identical to the one for a single independent sky pixel. And here also, no grouping or resampling of any kind is necessary such that all the information is used with full resolution to yield maximum accuracy in our estimates.

If the intensity measurements are derived from snapshots, as is often the case at many wavelengths, then a time series is made up of intensities from a single pixel in the images as a function of time. If instead, measurements consist of the detection of individual photons, then this can be treated either as in the general illustration of the method in §2.3, or, to use every event as each one is detected, the rate parameter can be estimated directly by the inverse of the time between events.

For example, a quarter of a second between two events gives a value of the estimate of the intensity in events per second, (the rate parameter), of 4 s^{-1}; if five seconds pass, then it is 0.2 s^{-1}. Therefore, whether the intensity is measured in a sky pixel as a function of time in successive images, whether it is the number of events detected during a given time interval, or whether it is estimated by the inverse of the inter-arrival times for each and every detected photon, the procedure remains the same in all respects and is as in §2.3.

Figure 2 is a demonstration of the technique applied to an unbinned time series (a list of event arrival times). In this example, we are observing a hypothetical flaring X-ray source embedded in a region from which the average event rate is one count per second. The observation lasted one hour, and in it occurred a weak flare that lasted about 30 seconds that counted exactly 33 events from start to finish. Even though a hint of its presence is seen in the 30 second binned time series (but not with 60 second bins), this event could easily have gone unnoticed if the detection relied on the examination of similar time series.

With the procedure described above, the burst is detected at a log-likelihood of -31.85, and thus likelihood of about 10^{-14}. This is the joint likelihood of detecting at least six consecutive events, each of which met the warning threshold, when expecting the average detected rate. In contrast, looking at the peak that stands out in the 30 second binned time series, we would compare a total intensity of 53 events against the expectation of 30, and find a likelihood of 5.8×10^{-4}, which might be enough to make us raise an eyebrow, but not much more. This also helps illustrate the important difference between the binned treatment of the data and the unbinned likelihood approach that treats each measurement individually.

In this example with a weak transient, the warning threshold was set at the relatively high value of 0.05 (1/20), but the detection trigger was set to require six consecutive events over the warning threshold.[4] Similar searches will require this kind

---

[4] This detection strategy was defined using simulations of observations that did not include a burst and ensuring that false transients were detected less than about 1% of the time, but that when a burst was included it was almost always detected.

of strategy. For detecting strong but short-lived transients, it would instead be better to use a very low, single-point threshold. Each application has its own purpose and thus optimal settings that can be very precisely tuned with simulations.

Related considerations will be discussed in detail in a subsequent paper, but it is worth noting again the significant difference between using the approach described here compared to that of binning the time series and looking for outliers. Although to establish the optimal warning threshold and the number of consecutive warnings required to define a transient is indeed akin to defining, by whatever means, the optimal bin time for a particular kind of transient (e.g. Scargle (1998)), it is very different in several ways. Using each measurement individually gives access to the full resolution of the data, and thus the exact distribution in time of the measurements. This, in turn, allows us to legitimately ask the question for each measurement 'what is the likelihood of this measurement if we expect that?' (the single-measurement likelihood), and also to ask the question 'what is the likelihood of these two, three, four or however many measurements taken together as an ensemble, when we expect this value?' (the joint likelihood), and the answer is always perfectly well-defined and directly interpretable as statistical evidence with respect to the question we are asking, (the hypothesis we are testing). Using each measurement also allows us to discriminate between consecutive measured values above the warning threshold, and an concurrence of several separate fluctuations above that threshold that are not consecutive but happened to occur in the same bin time interval.

3.4. *Transients in Energy Spectra*

In energy spectra, frequency histograms as a function of energy, transient features can also appear. When they do, they can potentially give powerful clues as to the physical changes in the conditions of the system under study. As with images, there are two ways of approaching the problem: each energy channel can be treated as independent of the others, in which case the procedure is the same as that for independent pixels, or we can recognise that the intensity in each spectral channel is almost surely related to the intensity in every other channel because radiative processes, other than those that manifest in mono-energetic line emission, produce a continuum of radiation. The former is simple, relies on no additional information and can thus be very powerful to detect narrow features. It is, however, limited in its sensitivity to changes that occur in several channels simultaneously. The latter is more complex and relies on the use of a global spectral model in order to treat the problem with all channels considered together, but is substantially more sensitive to subtle changes.

Hence, in the absence of a model, we use the former approach, treating each energy channel individually. We emphasise the importance of using all spectral information provided by the instrument and not grouping channels into coarser energy bins because *every measurement counts*, and each one must be used and accounted for.

Thus, with the very first measurement that falls in a single energy channel, the procedure is initiated and we have an estimate of the expected value that is used to construct the first likelihood function for that channel; with the second in the same channel, the likelihood ratio of that measured value with respect to the previous is calculated and evaluated for its potential as signalling a transient event, following which the joint and single-measurement likelihood functions are updated; with the third and every subsequent measurement, the same procedure is followed.

With a reliable model, even if it is simple, the power to detect changes in the shape of the spectrum increases markedly because the amount of information contained within and used to iteratively update the likelihood function is greater: All the measurements in each channel add to our knowledge of every other channel and makes the overall likelihood function as informative as it can be. The means by which it is calculated and updated is slightly different. In whichever way the measured values are distributed in a channel, it is the model that links these across the entire spectral range, and the key component from which the likelihood function is constructed and on which the entire procedure hinges.

From an initial estimate of the model's parameter values we define the joint likelihood function. In the case of Poisson distributed data, the function is the product of the contributing elements from each channel, each supplying a term of the form given in (A1):

$$f(n;\nu) = \frac{\nu^n e^{-\nu}}{n!}$$

This yields an ensemble of such terms, each with a different value of n: n_i—the measured intensity in channel i, and a different value of ν: ν_i—the model-predicted intensity in that channel. There is thus a single likelihood function: the joint likelihood function for the model across all spectral channels, and it is given by (A4):

$$L(\boldsymbol{\nu}|n) = \prod_i \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

Here as in the other cases, every single additional measurement brings an additional value of n_i in channel i from which the accuracy of estimates of ν_i can be improved. But unlike the procedure for independent pixels or spectral channels, the joint likelihood function is used to determine the most likely model parameter values given the available data, and this is done by fitting using the C statistic (Cash 1979) given by $C = -2 \ln L$, with $\ln L$ given by (A5). Since fitting entails varying the values of the parameters in order to maximise the likelihood and thus minimise the value of the fit statistic, comparing different model parameters is done through the ratio of likelihoods that translate to differences of log-likelihoods. Hence, terms (in this case the term) that does not depend on the parameters, will not contribute and can be dropped. The fit statistic becomes:

$$C = 2 \sum_i [\nu_i - n_i \ln(\nu_i)] \tag{1}$$

Thus what is updated with each measurement or iteration are the model's parameter values, and the identification of transient spectral features is based on the likelihood of the current freshly updated observed spectrum in relation to the previous best fit model. The thresholds are optimised for the application, both in terms of the individual value of the likelihood ratio after each measurement and in terms of the number of consecutive unlikely values. A discussion of the historical context, the motivation for, and the details that relate to procedures used in comparing data and models is the subject of another paper in preparation, and we therefore do not go any further into this here.

3.5. *Transients in power spectra*

The power spectrum is estimated by the periodogram and carries information about various scales in the observed system. Because it presents what is called 'power', which is a
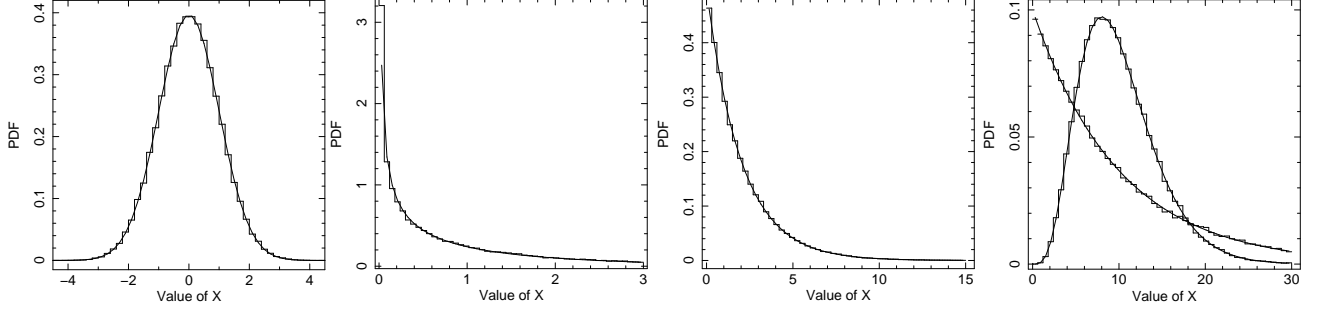
Figure 3. Relationship between standard Normal, χ^2 and Exponential variables
Making measurements of the value of a standard Normal random variable ($\mu = 0$, $\sigma^2 = 1$) and compiling the results into a histogram yields a distribution such as the one shown in panel 1. If the value of each measurement is squared, the result is a distribution such as the one in panel 2: it is a χ^2 of one degree of freedom or χ_1^2 distribution ($k = 1$: $\mu = 1$ and $\sigma^2 = 2$). If this process is repeated twice and the squares of the standard Normal variates are summed the result is as shown in panel 3 and yields a χ_2^2 distribution ($k = 2$: $\mu = 2$ and $\sigma^2 = 4$). If a χ_1^2 is summed ten times, or equivalently, if a χ_2^2 is summed five times, the resultant unimodal distribution peaking at a value of eight is as shown in panel 4: this is a χ_{10}^2 ($k = 10$: $\mu = 10$ and $\sigma^2 = 20$). Note that it is quite distinct from the distribution that results from multiplying or scaling a χ_1^2 by ten or χ_2^2 by five, which yields the Exponential distribution $f(x) = (1/10)\exp(-x/10)$ also shown in panel 4. Each histogram is built from 10^5 pseudo-random numbers, its area normalized to 1, and the normalised frequencies overlaid with the corresponding density function.

measure of the amount of 'activity' at a given frequency or time scale, the power spectrum conveys information about both dynamical and distance scales. Any spontaneous and stochastic, or triggered and stimulated event that leads to a reconfiguration of the dynamics of the system, be it local or global, will generally manifest itself in the power spectrum. How this will then translate into the periodogram for a given geometry and observing conditions depends upon how large, long lasting, and coherent the event, and surely most importantly, on the sensitivity of the instrument with which the system is being observed. Nevertheless, such transients can potentially be detected and identified as such, and we seek the best means to do so.

The approach most resembles the treatment of energy spectra in its details. For power spectra as well, the problem can be treated as though the values in each channel, in this case frequency channels, were independent of one another—something that is only true when the power spectrum is globally flat with equal power at all frequencies—or it can be treated with the use of a global model that prescribes how the values in the different channels are related. Note that in both cases, the model can be theoretical, or empirical and constructed from the successive iterations of measurements and refinement of the estimated spectrum by computations of the periodogram. Therefore, conceptually and procedurally, this problem is the same as for the energy spectrum.

There are two differences, however. One is superficial: that models for power spectra are generally fewer, largely phenomenological and often simpler than those used for energy spectra. The other is fundamental: that the values of power estimates in frequency channels are distributed neither as Poisson nor as Normal variables, but are instead related to χ^2 and Exponential distributions. The reason is that each power estimate is calculated from a sum of squared standard Normal variates.

For an ensemble of detected events, each with its own arrival time, the simplest way to calculate the power of the fundamental harmonic at a given frequency f, is to map each arrival time t_i to its phase ϕ_i within the periodic cycle that corresponds to that frequency ($p = 1/f$) and calculate the Rayleigh statistic (Leahy et al. 1983):

$$R^2 = 2N(C^2 + S^2) \tag{2}$$

where C and S are defined as:

$$C = \frac{1}{N}\sum_{i=1}^{N}\cos\phi_i \quad \text{and} \quad S = \frac{1}{N}\sum_{i=1}^{N}\sin\phi_i. \tag{3}$$

Firstly, the expectation or average value of the functions $\cos\phi$ and $\sin\phi$ is zero: $\langle\cos\phi\rangle = \langle\sin\phi\rangle = 0$. Therefore so are those for C and S: $\langle C \rangle = \langle S \rangle = 0$.

Secondly, the variances of $\cos\phi$ and $\sin\phi$ both equal one half: $\mathrm{V}[\cos\phi] = \mathrm{V}[\sin\phi] = 1/2$. Therefore those of C and S are a factor of N times smaller: $\mathrm{V}[C] = \mathrm{V}[S] = 1/2N$

Finally, since $\mathrm{V}[mX] = m^2\mathrm{V}[X]$, where m is a numerical constant, the scaled variables $c = \sqrt{2N}\cdot C$ and $s = \sqrt{2N}\cdot S$ have a variance of $\mathrm{V}[\sqrt{2N}\cdot C] = \mathrm{V}[\sqrt{2N}\cdot S] = 2N\cdot\mathrm{V}[C] = 1$, and are both standard Normal variables ($\mu = 0$, $\sigma^2 = 1$).

This implies that:

$$R^2 = c^2 + s^2 = 2NC^2 + 2NS^2 = 2N(C^2 + S^2)$$

is the sum of the squares of two standard Normal variables. Squaring a standard Normal yields a χ^2 variable with one degree of freedom. And summing χ^2 variables yields another χ^2 variable with a number of degrees of freedom that is the sum of the degrees of freedom of the variables being summed (illustration in Fig. 3). Therefore, the power being the sum of two χ_1^2 variables is χ_2^2 distributed with a mean of two, variance of four and standard deviation of two. This is convenient due to the simplicity of the χ^2 density function when k equals 2, which is a pure exponential:

$$\chi_2^2(x) = \frac{1}{2}e^{-x/2} \tag{4}$$

The caveat here is that this is only true if the power estimates at different frequencies are independent, which is only true for non-variable processes: the same kind that lead to a globally flat power spectrum with equal power at all frequencies. This is therefore an ideal case that should be treated as such, and not given more attention than it deserves. Nonetheless, it is instructive and illustrates how Normal, χ^2 and Exponential probability densities can be related.

A natural choice for a general χ^2 fit statistic is twice the negative of the log-likelihood of (A17), $K = -2\ln L$, and dropping the term that does not depend on the parameters k_i yields:

$$K = -2\sum_i\left[\left(\frac{k_i}{2} - 1\right)\ln x_i - \frac{k_i}{2}\ln 2 - \ln\Gamma\left(\frac{k_i}{2}\right)\right] \tag{5}$$

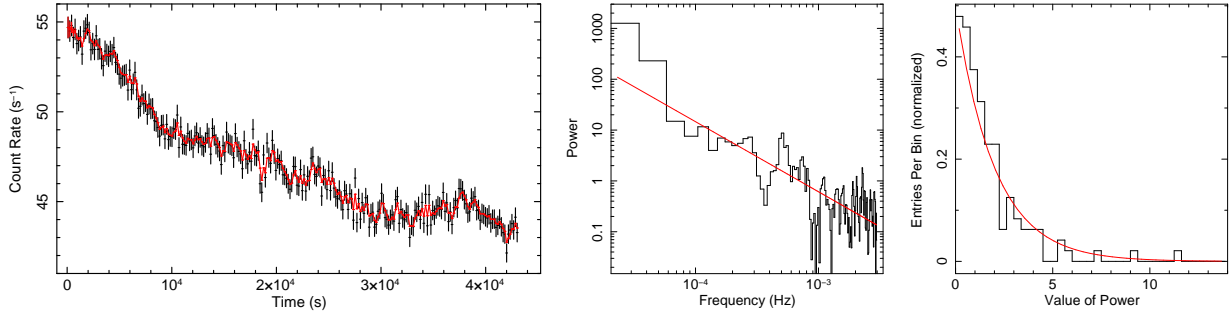**Figure 4. Periodogram powers of astrophysical red noise as scaled χ_2^2 variables**
Mkn 421 observed by XMM-Newton for 43 ks. Panel 1 shows the RGS time series in rates (0.3–2 keV with 85 s bins). In black are the measured data and in red are those that result from applying a Kalman filter (Kalman 1960) which very effectively suppresses the white noise scatter (see also König and Timmer 1997), and thus allows for a better constrained fit on the resultant periodogram that is shown in panel 2 with the best fit power-law model. Panel 3 is the distribution of de-trended periodogram powers overlaid with the analytical form of the χ_2^2 density function. The excess power at the lowest frequency, about 12 times above the best-fit in panel 2, is due to the global trend in the time series marked by a large difference between the start and end of the observation, and is seen in the right hand tail and greatest value in the histogram in panel 3.

Just as the C statistic is optimal for Poisson distributed data because it is derived from the likelihood function, which is in turn derived from the probability density for Poisson variables, the K statistic is optimal for χ^2 distributed data for the same reason. It is optimal for fitting a model to a set of measurements that are samples of random χ^2 variables with potentially different degrees of freedom k_i in each channel i.

Now, a much more common and also more general case than that of the flat spectrum of a non-variable process, is that of a power-law distribution of power estimates as a function of frequency usually called red noise in Astrophysics (see the highly sophisticated work by Vaughan (2010) on a Bayesian approach to evaluating the significance of an almost or strictly periodic signal in red noise for a discussion of its properties). The ideal case of a globally flat power spectrum considered above is the simplest manifestation of a red noise process: that with a spectral index of zero. The power values in red noise are related to one another in a well-defined manner through the relation $Power \propto f^{-\alpha}$, where f is the frequency and α is the power spectral index. This is a model that describes the shape of the underlying power spectrum; not the shape of any particular periodogram that serves as an estimate of it, subject to various degradation factors, distortions and statistical fluctuations.

As is the case for the likelihood, the absolute value of the power in the periodogram is not meaningful: it is only the relative power at a frequency compared to that at another that carries meaning. Therefore, any normalisation factor appropriate for the application can be used to scale the periodogram up or down. The key, however, is that we are working with *scaled χ_2^2 variables*. This is demonstrably true for astrophysical red noise (Fig. 4), because dividing the power estimates in each channel by the best-fit power-law model yields values that are χ_2^2 distributed.[5] This would also be true for *any* power spectral shape *if* the process can be considered as one of simply scaling the basic χ_2^2 variable that results from summing two squared standard Normal variables by the underlying model, whatever the shape. This is indeed what we have

always seen in our work in the frequency domain, and thus assume this to be generally the case.[6]

Therefore, whether we have a model describing the power spectral shape or not, we work with the power estimates at a given frequency as though they were χ_2^2 variables scaled differently in each channel. This implies they are distributed according to the Exponential density function (A20), that their joint likelihood function is given by (A23), and thus the log-likelihood by (A24):

$$\ln L(\boldsymbol{\tau}|x) = -\sum_i (\ln \tau_i + x_i/\tau_i)$$

In the periodogram, x_i is the measured, and τ_i is the model-predicted power in frequency channel i. We can therefore construct another fit statistic specifically for fitting periodograms based on this expression (Duvall and Harvey (1986) also derive and use this statistic) as was done above with K for the general case of different χ^2 variables, but now for the general case of a collection of exponentially distributed variables, such that $B = -2\ln L$:

$$B = 2\sum_i (\ln \tau_i + x_i/\tau_i) \qquad (6)$$

When working with a power spectral model, the B statistic is used to continuously compare the observed with the predicted distribution of powers in the periodogram, as is done with the C statistic in the case of Poisson distributed events in an energy spectrum, or the K statistic for χ^2 distributed measurements more generally. Sensitivity to detect changes in the overall shape of the spectrum increases quickly with the number of iterations. However, fluctuations in the power estimates in each channel always remains important due to the intrinsically high variance of Exponential variables for which the standard deviation is alway equal to the decay constant ($\mu = \tau$, $\sigma^2 = \tau^2$ and thus $\sigma = \tau$).

As a simple demonstration, we consider a hypothetical observation in X-rays of a bright (500 s^{-1}) accreting system whose variable emission comes mostly from two components: the accretion disk, and the hot and turbulent gas in the inner flow. In both, the emission processes are connected on all timescales, and thus each gives rise to red noise with a different spectral index. The accretion disk is much larger in extent

---

[5] It is important to demonstrate this using real data, because many algorithms used to generate power-law noise, as the standard one by Timmer and Koenig (1995), are precisely like this by construction, for the Fourier components are generated by multiplying the model spectrum at each frequency by pseudo-random standard Normal variates, one for the phase and one for the amplitude. Squaring and summing to get the power and then dividing by the best-fit power-law model will always give χ_2^2 distributed powers.

[6] Duvall and Harvey (1986) and Anderson et al. (1990) discuss this issue and also adopt this position.
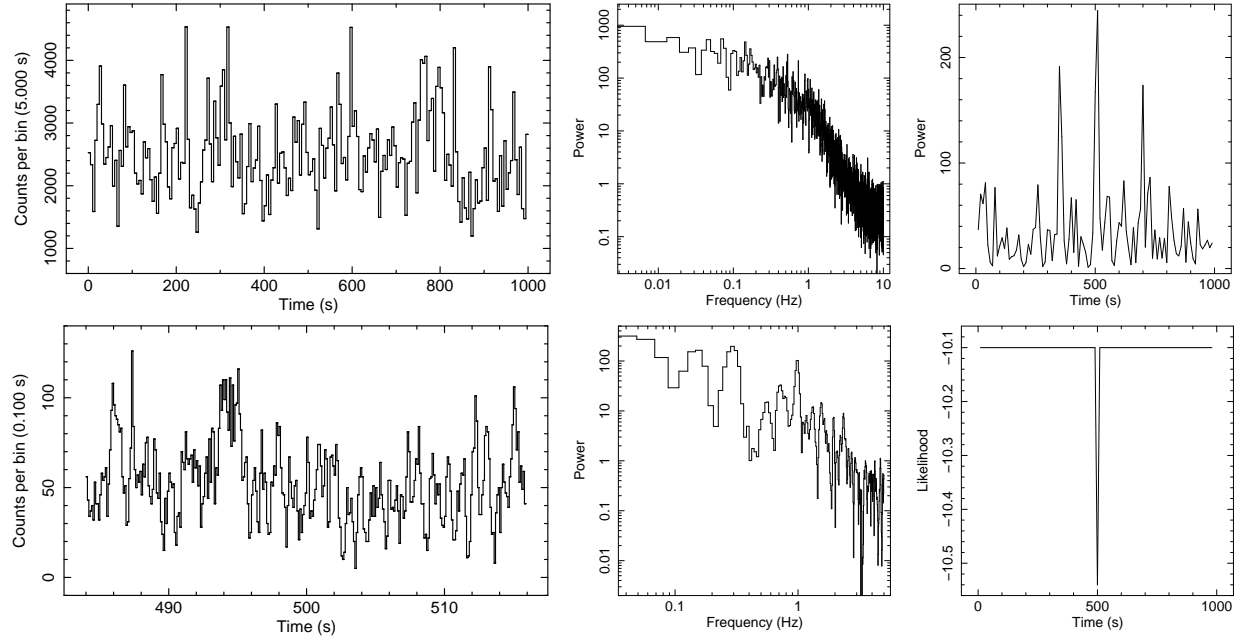
Figure 5. Detection of a transient feature in a periodogram
The top row shows the time series of the entire observation (binned to a resolution of 5 seconds for clarity of presentation); the periodogram made from the Kalman-filtered, 0.05 second resolution time series of the event arrival times; and the power at 1 Hz estimated at 10 second intervals from the Rayleigh periodogram of the event arrival times as a function of time. The bottom row shows a zoom on the time series during the transient QPO from its start at 485 seconds until its end at 515 second after the start of the observation; the periodogram of the kalman-filtered 0.05 second resolution time series; and the log-likelihood as a function of time where only detections beyond the established threshold are shown. The QPO is characterised by 30 cycles of an almost periodic signal centred on 1 Hz with a standard deviation of 1/20 about that frequency and a pulsed fraction of 27%.

and has a sharp inner radius. It dominates at lower frequencies with a power-law index $\alpha = -1$, and has a high-frequency cutoff beyond which it does not contribute to the power spectrum. The turbulent inner flow is much smaller in extent because it is bounded by the inner edge of the disk. Its emission is more variable and dominates the high-frequency part of the spectrum with a power-law index $\alpha = -3$.

In this case, we are interested in monitoring the range of frequencies between 0.1 and 10 Hz for the appearance of a weak, short-lived, transient QPO that we expect to appear at or very near the break in the power spectrum at 1 Hz, which marks the boundary between the disk and the turbulent inner flow. For this, we make a periodogram every 10 seconds with the events accumulated during this time interval, and monitor the power at one or any number of frequencies. Because we are interested in a short-lived QPO, the strategy is different than for the previous example of the time series: we cannot rely on the transient persisting in more than one "measurement", and therefore must establish a single detection threshold that is constraining enough for our application. This threshold is established using simulations. We have done this for the power at 1 Hz, the frequency of interest for us here, to first determine the average expected power (35), and then establish a threshold (log-likelihood of -10.1, and thus a likelihood of 4.1×10^{-5}) that ensures a level of false detections of 5%.

The observation and the analysis are presented in Figure 5 in which we see that the transient QPO is clearly detected in the likelihood monitoring, but, as should be expected from its very short lifetime, is not at all evident in the periodogram of the entire observations, even though it is indeed present. It is important to emphasise once more that to maximise the power of this method, the detection strategy must be devised and optimised for the application.

4. CONCLUDING REMARKS

Treating and interpreting data as statistical evidence in seeking to further our understanding of physical processes and of the behaviour of complex systems such as those we observe in astrophysics, using all the information carried by these data, is most directly done through the use of the likelihood function appropriately chosen according to the statistical distribution of the random variable that is measured. This approach yields a most powerful and effective means for treating any kind of data, because no information about the process under investigation is left out or ignored, nor degraded in any way; everything is seemlessly taken into consideration and into account in the statistical decision-making process. This is particularly appropriate for the search and identification of transient phenomena in all data domains of interest in astrophysics (temporal, spatial, energy and frequency), and this is what we have tried to demonstrate in this article.

The method presented is well suited to handle the first two classes of transients that have a non-variable background, be it nil or of some constant level, without any further considerations. Evidently, in this as in any other method, the identification efficiency ultimately depends very intimately on the relative strength of the transient signal. Identifying transients with the described procedures is perfectly suited for analysing archival data, where the data sets are already complete and well defined. It is, however, also powerful for real-time applications, maybe especially so. Handling the third class of transients characterised by a variable background requires additional care, as discussed in the opening sections (§2). Here are some additional considerations:

If the process is variable but predictable as with a sinusoidal modulation for example, then this is a simple extension of the procedure using a model, but in which it itself evolves as a

function of time; the formalism is otherwise identical. If the process is variable and unpredictable, it implies that the measurements in each pixel or channel are not distributed according to a particular and unchanging probability distribution. Instead, even though it may belong to a given family of distributions based on the nature of the process, it will inevitably have a hybrid shape due to the changes in that process as a function of time which we must model empirically. Therefore, each pixel or channel is treated individually, but because we have no a priori expression for the likelihood function, the intensity and how it is distributed can be characterised approximately using the running mean and variance.

For highly variable processes, where deviations in shape from known probability distributions are large, looking at the distribution of measured values per pixels or channel does not make much sense because the changing intensity in each is like a very small window onto complex interactions that give rise to variable emission that cannot be described by random variables with stationary probability distributions. Doing this is very limited in usefulness even when it is possible to find a function that can be fitted to the distribution of measured intensities such as a log-normal distribution, for example. However, a variable process can be highly non-stationary in the time domain, but stationary in frequency space, having a power spectrum that does change in time. This is analogous to a variable point source whose intensity varies markedly in successive images but whose location in the sky remains the same, and whose shape as it appears in each image is as always given by the instrument's PSF. Combining the information carried by the data in the time and frequency domains, and treating it simultaneously in the fashion described in this paper is certainly a most powerful means of analysis for detecting transient features in highly variable processes.

Note that, as alluded to in §2, the crucial elements in working with variable processes are the time scales involved: in this application, that of the transient with respect to that of the variability. More specifically, since the stationarity of the probability distribution can be considered as being a function of the time scale at which the process is viewed, it is in general possible to have a running estimation of that probability distribution which is stationary up to a given time scale, but evolves on longer time scales. In this way, the likelihood function and all the associated statistics are well defined at any point in time, and the method becomes a more general, time-dependent form of the procedure presented. As stated, details relating to this will be presented elsewhere.

The generality and wide applicability of the method we presented in this paper, can also be viewed as its primary limitation. This is not unlike the majority of statistical methods with a specific purpose, and is related, very simply in this case, to the fact that there are various kinds of transients, and we want to detect them all with maximum sensitivity. Therefore, as was shown in both the case of a transient in a time series and in a power spectrum, the power of the method relies on simulations for an accurate estimation of the statistics of the process, and for defining the detection thresholds, that will in general be geared towards a certain class of events. Nonetheless, there are in principle no limits to the number of automated searches that can be performed in parallel, for any given data set or application. And furthermore, the generality of the formalism is such that it can be applied to identifying transients in other parameter space, where the independent variable is not time.

Another limitation relates to the full reliance on well-defined statistics because the likelihood function can, other-wise, simply not be constructed. Although this may not appear to be an important factor in many methods commonly used, the truth is that it always is, but it is not necessarily recognised because of the generalised implicit assumption of Normal statistics. The periodogram is an excellent example of the importance of using the correct probability distribution.

Having recognised that the power values in a frequency channel of any periodogram are exponentially distributed with a mean given by the expected power in that channel, the one-sided tail probability of finding a power value of 60 or greater, for instance, when the expectation is 30, is 0.135 or 13.5%, which everyone would agree is definitely not worthy of even a second glance in terms of "statistical significance". However, using Normal statistics, (mean power of 30 and standard deviation of $\sqrt{30}$, say), finding a value of 60 or greater is a 5.47σ result with a probability of about 10^{-8}.

Therefore, although this limitation could be a weakness of the method in certain circumstances, it is clearly also a strength that highlights a very fundamental point which should rightly be incorporated in any statistical method: that of using the correct statistics.

We think this likelihood-based approach is a way forward, and will prove to be very valuable in many applications where transients and outliers are of importance.

APPENDIX

A. PROBABILITY DENSITY AND LIKELIHOOD FUNCTIONS OF POISSON, NORMAL, χ^2 AND EXPONENTIAL VARIABLES

A.1. *Poisson Variables*

The Poisson probability density function is given by:

$$f(n;\nu) = \frac{\nu^n e^{-\nu}}{n!} \qquad (A1)$$

where n is a random variable representing the number of detected counts during a specific time interval—a positive integer; and ν is the rate parameter that represents the number of counts expected in the same time interval—a positive real number. The likelihood function for a single measurement is given by the density function, as proportional to it in the sense described above:

$$L(\nu|n) \propto f(n;\nu) = \frac{\nu^n e^{-\nu}}{n!} \qquad (A2)$$

The semi-colon is used to separate n and ν to indicate that ν is a parameter and not a variable, and that its value must be fixed in order to know the form of a specific density function. It is not a conditional probability of a bivariate Normal density function, for example, where we fix the value of y and look at the density function for x, and is written $f(x|y)$. To write the likelihood function, we use $L(\nu|n)$ to indicate that it is a function of the parameter ν given the measured data that is now fixed. We generally use '=' instead '∝' for simplicity.

For more than one measurement, a vector of measurements n where each is now identified by the subscript i, the likelihood function is:

$$L(\nu|n) = \prod_i \frac{\nu^{n_i} e^{-\nu}}{n_i!} \tag{A3}$$

If we are instead working with a model and comparing the measured number of events with what the model predicts in each channel of a spectrum, for example, then each measurement n_i is compared with the model-predicted parameter ν_i and the expression for the joint likelihood becomes:

$$L(\boldsymbol{\nu}|n) = \prod_i \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!} \tag{A4}$$

It is often convenient to work with the natural logarithm of the likelihood function in which products become sums, ratios become differences, and powers become products.

$$\ln L(\boldsymbol{\nu}|n) = \sum_i [n_i \ln(\nu_i) - \nu_i - \ln(n_i!)] \tag{A5}$$

A.2. *Normal Variables*

The univariate Normal probability density function is:

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{A6}$$

where the variable x is a real number, μ is the mean and σ the standard deviation. There are now two parameters, μ and σ, instead of a single one as for the Poisson distribution. Because the two parameters are independent, the likelihood function can be expressed for either one of them (fixing the other) or for both, in which case, for a single measurement, it is:

$$L(\mu,\sigma|x) \propto g(x;\mu,\sigma) = e^{-(x-\mu)^2/2\sigma^2} \tag{A7}$$

and for a collection of measurements it becomes:

$$L(\mu,\sigma|x) = \prod_i e^{-(x_i-\mu)^2/2\sigma^2} \tag{A8}$$

If we are comparing each measurement with a model-predicted value, the likelihood function is:

$$L(\boldsymbol{\mu},\boldsymbol{\sigma}|x) = \prod_i e^{-(x_i-\mu_i)^2/2\sigma_i^2} \tag{A9}$$

and its log-likelihood is given by:

$$\ln L(\boldsymbol{\mu},\boldsymbol{\sigma}|x) = -\sum_i (x_i-\mu_i)^2/2\sigma_i^2 \tag{A10}$$

A.3. *Pearson Type III or χ^2 Variables*

The χ^2 probability density function is given by:

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{A11}$$

where the variable x, is equal or greater than zero and continuous on the real line, and the parameter k, called the number of degrees of freedom of the distribution, is traditionally defined to be a positive integer, but can be any positive real number (how it is used here). As is the case for the Poisson distribution, the parameter determines the shape of the curve.

The Γ function is a generalised factorial defined as:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \tag{A12}$$

and has the following relations:

$$\Gamma(x+1) = x\Gamma(x), \quad \Gamma(n) = (n-1)! \quad \text{and} \quad \Gamma(1/2) = \sqrt{\pi}. \tag{A13}$$

The single-measurement likelihood function is therefore:

$$L(k|x) \propto f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{A14}$$

and the joint likelihood for multiple measurements is:

$$L(k|x) = \prod_i \frac{1}{2^{k/2}\Gamma(k/2)} x_i^{k/2-1} e^{-x_i/2} \tag{A15}$$

For the general case of a collection of samples drawn from different χ^2 distributions, each with a given value of k, the joint likelihood function is given by the generalisation of (A15) where k is replaced by k_i:

$$L(k|x) = \prod_i \frac{1}{2^{k_i/2}\Gamma(k_i/2)} x_i^{k_i/2-1} e^{-x_i/2} \tag{A16}$$

Taking the natural logarithm and simplifying the expression using the properties of the logarithm yields the log-likelihood:

$$\ln L(k|x) = \sum_i \left[\left(\frac{k_i}{2} - 1 \right) \ln x_i - \frac{x_i}{2} - \frac{k_i}{2} \ln 2 - \ln \Gamma \left(\frac{k_i}{2} \right) \right] \tag{A17}$$

where $\ln \Gamma(z)$ is available in most programming environments, but can also be approximated by:

$$\ln \Gamma(z) \approx (z-1/2)\ln z - z + \frac{1}{2}\ln(2\pi) \tag{A18}$$

Substituting $z = k_i/2$ gives:

$$\ln \Gamma \left(\frac{k_i}{2} \right) \approx \frac{1}{2}(k_i-1)\ln\left(\frac{k_i}{2} \right) - \frac{k_i}{2} + \frac{1}{2}\ln(2\pi) \tag{A19}$$

In all equations, k_i is the model-predicted numbers of degrees of freedom in channel i, and x_i is the value of the random variable measured in that channel. This implies that there is for each channel a different value of the parameter k and a single measurement x. If there are several measurements for each channel, then each one would have a joint-likelihood function $L(k_i|x)$ resulting from the product of j measurements. The overall joint-likelihood would then be the product of these joint-likelihoods per channel.

The χ^2 distribution is a single-parameter function: its mean is the number of degrees of freedom, and the variance is twice that. This implies a large scatter about the mean that is proportional to it in amplitude. The mode—the most likely value where the probability peaks—is not the mean as for the Normal and Poisson distributions; it is given by $\max(k-2, 0)$.

A.4. *Exponential Variables*

The Exponential probability density is given by:

$$f(x;\tau) = \frac{1}{\tau} e^{-x/\tau} \tag{A20}$$

where x is the random variable—a real number greater or equal to zero, and τ—a real number greater than zero, is called the decay constant because it defines the speed at which the function decays. The mean of the distribution equals τ

and its variance equals τ^2. The single-measurement likelihood function is:

$$L(\tau|x) \propto f(x;\tau) = \frac{1}{\tau}e^{-x/\tau} \tag{A21}$$

The joint-likelihood of a set of measurements drawn from the same distribution (same τ, same frequency channel) is the product of the individual probabilities:

$$L(\tau) = \prod_i \frac{1}{\tau}e^{-x_i/\tau} \tag{A22}$$

The joint-likelihood for a set of measurements—one per channel—across the entire spectrum is:

$$L(\boldsymbol{\tau}) = \prod_i \frac{1}{\tau_i}e^{-x_i/\tau_i} \tag{A23}$$

And the corresponding log-likelihood function is given by:

$$\ln L(\boldsymbol{\tau}) = -\sum_i (\ln \tau_i + x_i/\tau_i) \tag{A24}$$

REFERENCES

R. A. Fisher, Messenger of Mathematics **41**, 155 (1912).
R. A. Fisher, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **222**, 309 (1922).
R. M. Royall, *Statistical Evidence*, Likelihood Paradigm (Chapman & Hall/CRC, 1997).
G. Cowan, *Statistical data analysis* (Statistical data analysis. Publisher: Oxford: Clarendon Press, 1997).
J. D. Scargle, Astrophysical Journal v.504 **504**, 405 (1998).
J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang, The Astrophysical Journal **764**, 167 (2013).
M. Bayes and M. Price, Philosophical Transactions of the Royal Society of London Series B Containing Papers of a Biological Character (1896-1934) **53**, 370 (1763).
W. Cash, Astrophysical Journal **228**, 939 (1979).
D. A. Leahy, R. F. Elsner, and M. C. Weisskopf, Astrophysical Journal **272**, 256 (1983).
R. E. Kalman, Journal of basic Engineering **82**, 35 (1960).
M. Konig and J. Timmer, A & A Supplement series **124**, 589 (1997).
S. Vaughan, Monthly Notices of the Royal Astronomical Society **402**, 307 (2010).
J. Timmer and M. Koenig, Astronomy and Astrophysics **300**, 707 (1995).
T. L. J. Duvall and J. W. Harvey, in *IN: Seismology of the sun and the distant stars; Proceedings of the NATO Advanced Research Workshop* (NASA, Goddard Space Flight Center, Greenbelt, MD, 1986), pp. 105–116.
E. R. Anderson, T. L. J. Duvall, and S. M. Jefferies, Astrophysical Journal **364**, 699 (1990).